****

**BAHIR DAR UNIVERSITY**

**BAHIR DAR INSTITUTE OF TECHNOLOGY**

**FACULITY OF COMPUTING**

**Course Title: Natural Language Processing**

**Review On**

**Develop Corpus Based IR System for Amharic Language Using Vector Space Model**

**Name: Msge Desta Akalu**

**ID no: BDU1500526**

**Submitted to**: Mr. Alemu K. (Assistant prof.)

Date: February, 2019

**Acknowledge**

First of all, I would like to thanks to **Mr. Alemu** for giving this mini-project which assist me as I get more knowledge from it by doing practically about how I can process natural language tasks and use it for research works.

Table Contents

[**Acknowledge** ii](#_Toc128465063)

[List of Figures iv](#_Toc128465064)

[Abstract v](#_Toc128465065)

[1 Introduction 1](#_Toc128465066)

[2 Problem Statement 2](#_Toc128465067)

[3 Objective of the project 2](#_Toc128465068)

[3.1 General objectives 2](#_Toc128465069)

[3.2 Specific objective 2](#_Toc128465070)

[4 General process in IR System 3](#_Toc128465071)

[4.1 Indexing Subsystem Structure 3](#_Toc128465072)

[4.2 Searching Sub System 5](#_Toc128465073)

[5 IR MODELS 6](#_Toc128465074)

[6 Methodology 8](#_Toc128465075)

[6.1 Literature review 8](#_Toc128465076)

[6.2 Dataset Preparation 8](#_Toc128465077)

[6.3 Description of architecture of IR system 9](#_Toc128465078)

[6.4 Preparing stop list 9](#_Toc128465079)

[6.5 Stemming words to their original root 9](#_Toc128465080)

[6.6 Indexing content bearing words only 9](#_Toc128465081)

[6.7 Preparing a vector space of documents and indexing terms 9](#_Toc128465082)

[6.8 Calculate the similarity, finding documents, ranking and display documents 10](#_Toc128465083)

[7 Discussion of evaluation result, with findings 10](#_Toc128465084)

[7.1 Implementation Tools 10](#_Toc128465085)

[7.2 Experimentation 11](#_Toc128465086)

[7.3 Experimental result 11](#_Toc128465087)

[8 Conclusion and Recommendation 13](#_Toc128465088)

[9 Reference 14](#_Toc128465089)

# List of Figures

[Figure 1 General Process of IR system 3](#_Toc128464752)

[Figure 2Indexing subsystem structure 5](#_Toc128464753)

[Figure 3 Searching Subsystem 6](#_Toc128464754)

# Abstract

Information retrieval (IR) is a mechanism that enables finding relevant information material of unstructured nature that satisfies information needs of user from large collection. Since there are usually many ways to express the same concepts, the terms in the user’s query may not appear in a relevant document. Alternatively, many words can also have more than one meaning which may confuse the retrieval system.

The Web is a huge repository of information in the form of text, image, audio, and video. People use search engines, such as Google, Yahoo!, Bing, etc. to discover resources from this huge repository. Originally these general-purpose search engines are designed and optimized for English language. They fell short when they are used for locating web resources of other languages such as Amharic. This is mainly due to the specific features of the language that are not considered by those search engines. Amharic, which is a family of Semitic languages, is the working language of the federal government of Ethiopia. Currently, there are significant numbers of Amharic documents on the Web. In this work, we analyzed the specific features of Amharic language, designed a general architecture for Amharic Search Engine, developed the necessary algorithm to realize it and implemented the same for searching Amharic language web documents. The result of the work has become to be a complete language specific search engine that has a crawler, an indexer and a query engine component that are optimized for the language they are designed, Amharic language.

# Introduction

Now a days Information Retrieval (IR) plays a vital role in human daily life through its implementation in a range of practical applications, including searching the web, question answering systems, personal assistants, chatbots, and digital libraries among others(Hambarde & Proenca, 2023). The primary objective of IR is to locate and retrieve information that is relevant to a user’s query. The main goal of IR is to identify and retrieve information that is related to a user’s query. As multiple records may be relevant, the results are often ranked according to their relevance score to the user’s query. Several techniques including a simple keyword to advanced NLP are available for developing IR systems. Various available IR models include a Boolean model, Vector Space Model (VSM), Probabilistic Model, Language Model.

In developing IR system stemming process is performed in improving performance if the system. Stemming serves two purposes in the context of Information Retrieval (IR): (i) reducing the index size and (ii) recall enhancement in the document lists retrieved for queries. Mapping actual query terms to their base forms reduces the number of unique terms in any given collection and as a result it proportionally reduces the size of the index(Göksel et al., 2023).

This mini-project, an Amharic Language based IR system stemming, is designed to show the challenges of information retrieval regarding text preprocessing, preparation of inverted index files, posting files, term weighting, similarity measure and ranking.

In this system the IR process involves two major sub parts; indexing and searching. Indexing is a process of creating logical representation of documents that are found in the corpus. It is an offline process of organizing and representing large document collection using index term. On the other hand, Searching is the online process that maps users’ information need (represented in the form of query) with the documents representation (index files) by using matching methods and returning relevant documents from the collection of unstructured or semi-structured corpus to satisfy users need.

1. Problem Statement

The traditional text retrieval systems mainly rely on the matching of terms between the

query and the documents. However, term-based retrieval systems have several limitations such as polysemy, synonymy, and lexical gaps between the query and the documents. Finding useful information from large amount of stored document is the most tedious and difficult task. For instance, to satisfy the information need of the user, the user might navigate the space of Web links (i.e., the hyperspace) searching for information of interest. There are lots of challenges we faced while we are retrieving information

The main problems happen while retrieving information from the source is

* **Difficulty in locating relevant document as per user requirement**
* **Time consumption:** when we are going to get the required document, we will consume a lot of time to get it.
* **Manpower consumption:** and also, while we are struggling for finding the relevant information, we will lose our selves
* **User Unsatisfaction of information need**

So, to overcome this problem a well designed and implemented Information retrieval system using deep learning algorithm like CNN, RNN and transfer learning is required.

1. Objective of the project
   1. General objectives

The general objective of the project is design and develop corpus-based IR system for Amharic language using vector space model.

* 1. Specific objective

The IR system designed incorporates the following tasks

* To prepare a corpus for the IR system
* Data preprocessing and analysis like tokenization, normalization, removing stop words, and stemming
* Indexing content bearing text
* Implementing system that prepare Inverted file that contains is vocabulary (D#,Df, Cf)and posting(T,Tf, location)
* Term weighting measuring using TF\*IDF
* Similarity measure among query and documents of the corpus using cosine similarity.
* Ranking documents on their relevance.

# General process in IR System

There are three primary processes an IR model has to support: the representation of the information of the documents, the interpretation of the user's information need, and the comparison of these two representations(Saini et al., 2014).

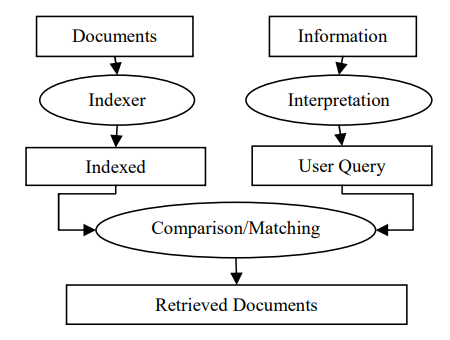


Figure 1 General Process of IR system

## Indexing Subsystem Structure

The indexing system generate index terms for a given Amharic text document and organize them using inverted file indexing, include TF, DF & CF for each index term and position/location information of terms in each document.

There are various common information retrieval indexing techniques. These are

* Signature files and
* Inverted index

**Signature file:**

In signature file indexing technique, each document returns a bit of string, (that is, signature) using hashing method on its text and superimposed coding. The final output of document signatures are stored in a special way, that is sequentially in a separate file and this file is called as signature file. The signature file is much smaller than the original file, and it can provide high search rate.

**Inverted index:**

Each document can be represented by a list of some reference words called keywords which depict the contents of the document for retrieval purpose. Fast retrieval can be obtained if we invert on those keywords. All the reference words are stored alphabetically in a file called index file. For each keyword we keep a list of pointers to the characterize documents in the postings file. This method is mostly used by all the commercial systems.

Before indexing applying the following text pre-processing tasks will be executed for the document’s corpus:

* The text of each document of the corpus is taken
* Tokenization is performed by removing punctuation marks
* Stop words are removed. Here, words which are considered stop words are rejected. No statistical technique of distinguishing stop words is used to do so.
* The tokens which are content bearing are stemmed to their root word. But terms less than three characters long will not be stemmed. The stemming part takes longer time.
* The stemmed words are normalized and indexed

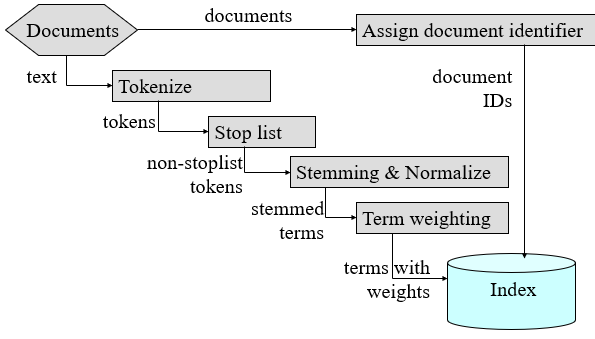
After executing all the above text operations, the index system will be built. The indexed system that used in this project is an inverted index file system. In the system both indexing (vocabulary) file and posting files will be prepared. 

Figure 2Indexing subsystem structure

## Searching Sub System

* The end user is requested for his/her query term/phrase. In the system, the query term/phrase is expected to be longer than one character long on the mere fact of poor content description of a character. And should be maximum of 14 words length.
* The input query is tokenized and Stop words are rejected from the query;
* Stemming is applied on the non-stop word terms;
* Normalization is accomplished. In Amharic language there are lots of variants for a letter. For example, assume the word in the document is read as “ባህርዳር”. If someone gives “ባሕርዳር” as a query.
* Vector space similarity measure is applied for retrieval of relevant documents
* The searching subsystem accepts a minimum length of user query 2 characters and a maximum of 14 words. If the user provides without this range his/her searching will not be successful.
* If the search result is not empty, the search system will ask the users interest to read any from the ranked documents, if interested Accept the document name and the system will display the content If not interested will ask to perform another search if interested
* Repeat the search process else stop the system.

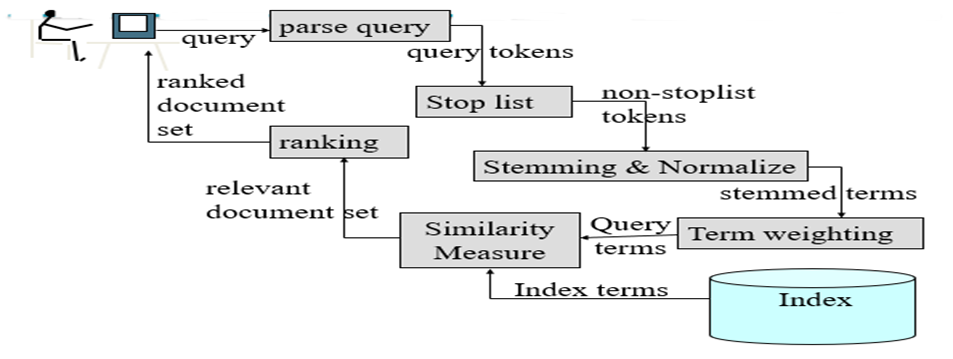


Figure 3 Searching Subsystem

# IR MODELS

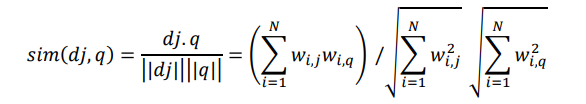
An IR model describes the elaboration of the document representation, the user query representation and the retrieval mechanism or process. By and large, three classic framework models have been used in the process of retrieving information: Boolean, Vector Space and Probabilistic.

1. **Boolean model:**

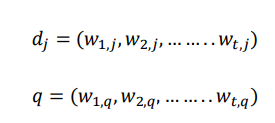
The Boolean model uses set theory, that is, Boolean algebra and its three components AND, OR and NOT, for query formulation, but it has one major drawback: a Boolean system is fail to rank the result list of retrieved documents. In the Boolean model, all documents are associated with a set of distinct words or key-words and User Queries are also represented by expressions of keywords separated by AND, OR, or NOT. The retrieval function of Boolean model takes a document as either relevant or irrelevant

1. **Vector-Space model:**

The Vector Space Model works on the concept of similarity. It makes an assumption that relevance between a document and the query is directly related to the similarity between their vector representations. In the Vector Space Model (VSM), documents and user query are both represented as a Vector and the angle between the two vectors are calculated using special function, that is, cosine function. Cosine function defines the similarity values between two given vectors and it can also be defined as: Where,



Documents and queries are represented as vectors.



Vector Space Model have been introducing term-weight scheme known as tf-idf weighting. These weights have a term frequency (tf) factor measuring the frequency of occurrence of the terms in the document or query texts and an inverse document frequency (idf) factor measuring the inverse of the number of documents that contain a query or document term.

1. **Probabilistic model:**

The most crucial or essential function of the probabilistic model is its initiate to rank documents by their probability of relevance given a user's query [5]. Both documents and user queries are represented by vectors ~d and ~q these are binary vectors, each binary vector component show whether a given document attribute/component or term occurs in the document or query, or not.

# Methodology

The following methods or techniques were used in order do the project and to achieve the objective of this project. Python 3.9 were used to develop the codes used to construct the text operation, indexing and searching system.

## Literature review

In order to understand the principles/theories and techniques of information retrieval system, particularly on Amharic language, extensive literature review has been conducted from books, articles and Internet.

Researcher (Manwar et al., n.d.), implements and discusses the issues of information retrieval system with vector space model using MATLAB on Cranfield data collection of aerodynamics domain and states that Inter-document characterization and document frequency plays vital role in building ranks of the documents in vector space model.

According to (Ahmad Abu-Salih, n.d.) implementation for a core IR technique which is Vector Space Model (VSM). They chosen VSM model for their project since it is a term weighting scheme, and the retrieved documents could be sorted according to their relevancy degree. One other significant feature for such technique is the ability to get relevance feedback from the users of the system; users can judge whether the retrieved document is relative to their need or not. The developed system has been validated through building an Arabic IR website using server-side scripting. The experiments verify the effectiveness of our system to apply all techniques of vector space model and valid over Arabic language.

## Dataset Preparation

In this project, Amharic documents collected from Amharic local news articles available on web. I use news articles as a test data because, news articles are easier to access, available in electronic form on the web. The following documents have been prepared from Amharic local news articles. Text corpus\doc1.txt, Text corpus\doc2.txt, Text corpus\doc3.txt, …., Text corpus\doc20.txt.

## Description of architecture of IR system

After developing the complete python code for the searching system, I prepare twenty Amharic text files that contain the Amharic texts that is going to be searched or serve as document corpus. By running the developed search system different time with the provision of different Amharic search queries and checked for accuracy of the search output.

## Preparing stop list

Due to shortage and unavailability of related documentation on Amharic stop lists, I was forced to make a decision on making my own stop list. So, I collected a number of Amharic document and make through and recursive search looking for stop lists that satisfy the judgment. After collecting these words, I put them in a file.

## Stemming words to their original root

All languages extend words by using prefixes, suffixes and infixes. Having this information, I tried to gather basic prefixes and suffixes of Amharic language. Having the suffix and the prefix lists, I remove them from a term that contains it.

## Indexing content bearing words only

I have implemented this by using a python built in object “re‟ to substitute all punctuation marks in Amharic by a space. After removing punctuation marks the next job is to check weather words are stop words or content bearing words. I have implemented this by copying all the stop lists I have stored in a file into a dictionary and checking each word found in each document against this list. So those words found in both will be removed from the index. The rest of the words that are not found in the stop list will be stored in the index file with their frequency of occurrence and the document it is found in.

## Preparing a vector space of documents and indexing terms

This project is implemented by using vector space model. In a vector space model, I have a multidimensional space that holds a relationship between terms and each document. For each term in the index file information about its frequency in each document, the total document frequency (dfi), the corpus to document frequency ratio for the term (D/dfi), the inverse document frequency (idfi), and the weight of each document in relation to each term will be stored.

## Calculate the similarity, finding documents, ranking and display documents

By using the cosine similarity formula, the distance between each document and the given query is calculated. The highest cosine similarity value will be more relevant to the query. After calculating these values for each document, I stored them in a dictionary data structure. The next step is ranking documents according to their level of similarity to a given query. I have implemented this by using descending sort algorithm.

Finally, after finding out the most similar documents with a threshold value of greater than zero and ranked them the only job left to be done was displaying the result to the user in Amharic language as to the scope of the project.

# Discussion of evaluation result, with findings

The Vector Space Model is a simple model based on linear algebra which was designed to overcome the limitations of the Boolean Model. One of the major advantages of VSM over the Boolean Model is that the weights assigned to the term are not binary.

During the different run times of the search system, different Amharic search queries was given to the system to search any relevant document. It was found that the system could retrieve relevant Amharic document if the search query was included in the document corpus and displayed the appropriate search result, and for those search queries that were not included in the document corpus the system did display a message indicating that it could not find any relevant document.

## Implementation Tools

I have used Python as a programming language for developing this system. Because of the reason the syntax is very clear and readable more user friendly, and more powerful language into days world I have used python programming language.

## Experimentation

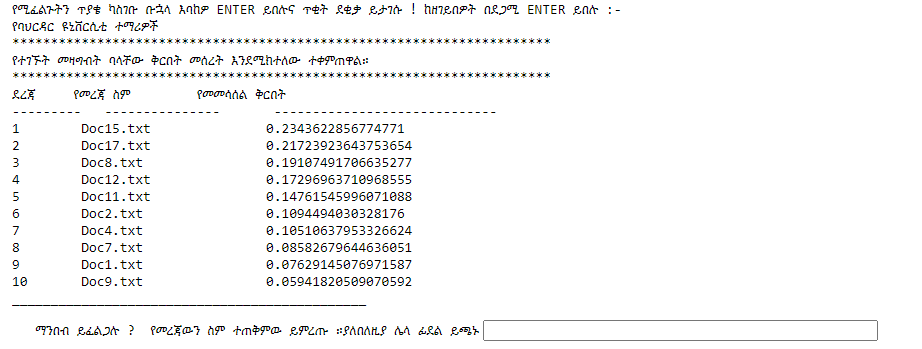
In this project we have design corpus-based IR system for Amharic language. The system has both the indexing and searching parts. Inverted file indexing structure is used to organize documents so as to speed up searching.

To test the developed system, sample texts were used as a document corpus. The corpus contains twenty documents. Each document articles are saved using .txt format, which is supported by python programming language. Then, test queries were selected to test the system. For retrieving the document which is relevant for a given query.

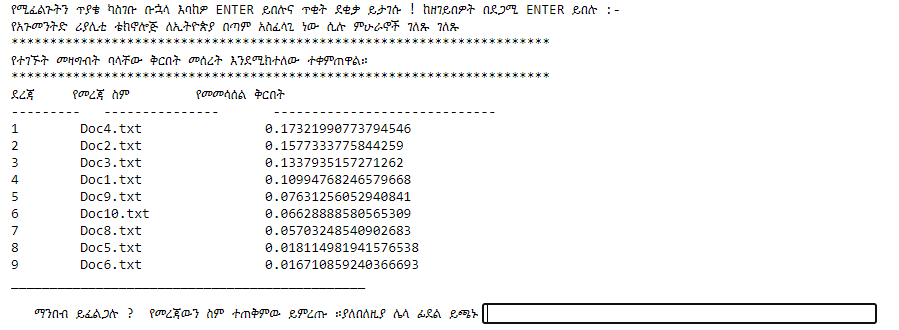
## Experimental result

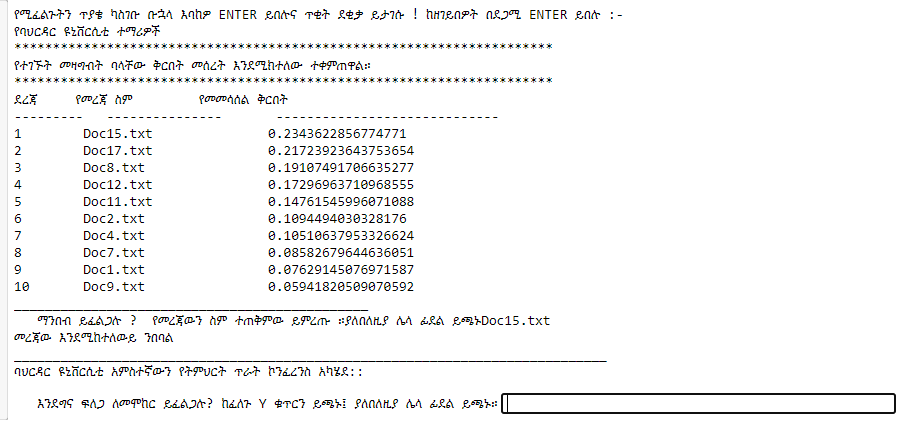
Searching: - the system is tested by inserting the following query with three, five and ten words and we have got the following results.

**Query with three words**

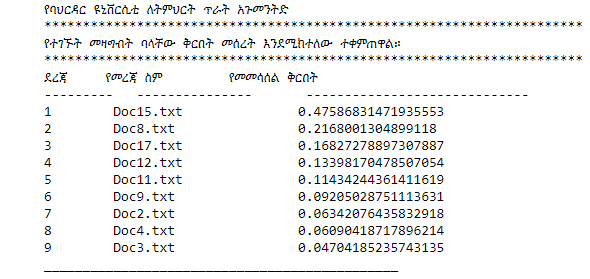


**Query with ten words**

The main evaluation measures to check the performance of an IR system by using Efficiency and Effectiveness. To know the precision and recall of the Query with three words it takes a little time approximately two minutes consume at the time of indexing and searching Amharic IR system.



To know the precision and recall of the Query with five words



# Conclusion and Recommendation

Even if developing a search system for Amharic document collection is a challenging work because of the unique nature of the language, this project indicated that it is possible to develop the search system as long as the Amharic document collection is represented in Unicode format.

I recommend that, in developing the search system, the great challenge was to create a power stemming system and this was extremely difficult as Amharic language has no uniform rule related to word stemming. Therefore, through study should be conducted to solve this problem of Amharic word stemming to get high performance.

# Reference

1. Vechtomova, O., *Introduction to Information Retrieval Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (Stanford University, Yahoo! Research, and University of Stuttgart) Cambridge: Cambridge University Press, 2008, xxi+ 482 pp; hardbound, ISBN 978-0-521-86571-5, $60.00*. 2009, MIT Press.

2. Alemayehu, N. and P. Willett, *Stemming of Amharic words for information retrieval.* Literary and Linguistic computing, 2002. **17**(1): p. 1-17.